

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-259487

(43)Date of publication of application : 24.09.1999

(51)Int.Cl.

G06F 17/30

(21)Application number : 10-055560

(71)Applicant : TOSHIBA CORP
TOSHIBA COMPUT ENG CORP

(22)Date of filing : 06.03.1998

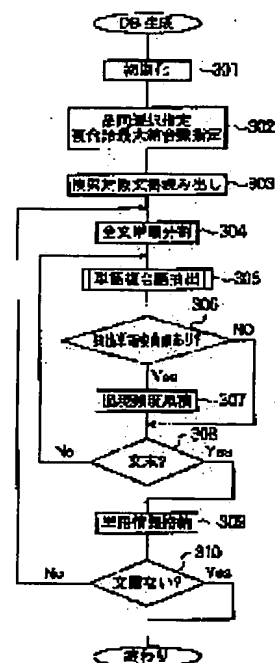
(72)Inventor : TANOSAKI YASUO
NISHINA TAKUYA
NAKAMOTO YUKIO
KUBOTA NAOHIDE

(54) SIMILAR DOCUMENT RETRIEVING DEVICE, SIMILAR DOCUMENT RETRIEVING METHOD AND STORAGE MEDIUM RECORDED WITH PROGRAM FOR RETRIEVING SIMILAR DOCUMENT

(57)Abstract:

PROBLEM TO BE SOLVED: To enhance retrieving accuracy when a similar document is retrieved by extracting a composite word from a retrieving key document and a document to be retrieved.

SOLUTION: When the composite word, i.e., 'brush character address printing function' exists in the retrieving key document or the document to be retrieved and the maximum number of connections is specified as three, all the composite words like 'brush character address', 'character address printing', 'address printing function', 'brush character', 'character address', etc., consisting of the number of words to be equal to or less than the maximum number of connections are extracted from each document, appearance frequency of the composite words is calculated and a degree of similarity between the retrieving key document and the document to be retrieved is calculated. Since different composite words to characterize the documents with specified contents are thoroughly extracted, a more proper degree of similarity between the documents is calculated and highly accurate retrieval of similar document intended by a user is performed.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision
of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japan Patent Office

W81

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-259487

(43) 公開日 平成11年(1999) 9月24日

(51) Int.Cl.⁶

G 0 6 F 17/30

識別記号

- F I

G 0 6 F 15/403

15/401

15/403

3 3 0 B

3 1 0 A

3 5 0 C

審査請求 未請求 請求項の数 5 O L (全 12 頁)

(21) 出願番号 特願平10-55560

(22) 出願日 平成10年(1998) 3月6日

(71) 出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(71) 出願人 000221052

東芝コンピュータエンジニアリング株式会社

東京都青梅市新町3丁目3番地の1

(72) 発明者 田野崎 康雄

東京都青梅市末広町2丁目9番地 株式会社東芝青梅工場内

(74) 代理人 弁理士 須山 佐一

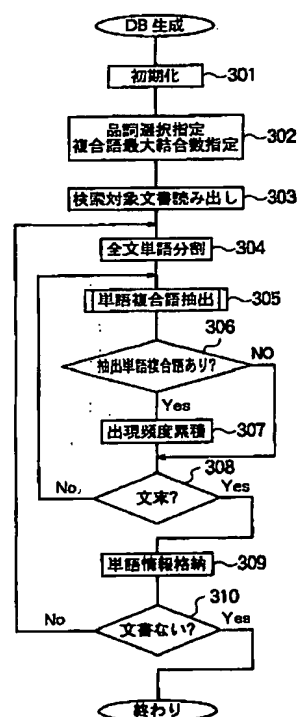
最終頁に続く

(54) 【発明の名称】 類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体

(57) 【要約】

【課題】 検索キー文書と検索対象文書から複合語を抽出して類似文書を検索する場合の検索精度の向上を図る。

【解決手段】 検索キー文書或いは検索対象文書に例えば「筆文字宛名印刷機能」といった複合語が存在し、最大結合数を3と指定したとき「筆文字宛名」「文字宛名印刷」「宛名印刷機能」「筆文字」「文字宛名」などの最大結合数以下の単語数からなる複合語を各文書からすべて抽出し、これらの複合語の出現頻度を計算して、検索キー文書と検索対象文書との類似度を算出する。特定の内容の文書の特徴付ける異なる複合語を漏れなく抽出することができるので、文書間のより妥当な類似度を計算でき、ユーザの意図する精度の高い類似文書検索を行うことができる。



【特許請求の範囲】

【請求項1】 ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、
前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、
前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語数の上限値を指定する指定手段と、
前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の数の単語により構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、
前記複合語抽出手段によって抽出された複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備することを特徴とする類似文書検索装置。

【請求項2】 ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、
前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、
前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語の数の上限値を指定する指定手段と、
前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の数の単語により構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、
任意の単語を不要語として選択する不要語選択手段と、
前記複合語抽出手段によって抽出された複合語のうち前記不要語選択手段によって選択された不要語を含む複合語を無効とする複合語無効化手段と、
前記複合語抽出手段によって抽出された有効な複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備することを特徴とする類似文書検索装置。

【請求項3】 請求項1または2記載の類似文書検索装置において、
前記単語抽出手段による単語の抽出条件として単語の品詞を指定する手段を有することを特徴とする類似文書検索装置。

【請求項4】 ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から

検索する類似文書検索方法において、
前記検索キー文書および前記検索対象文書を単語単位に分割し、
前記分割された単語の中から予め指定された条件を満たす単語を抽出し、
前記抽出された単語の結合により構成される複合語のうち予め指定された数以下の単語で構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出し、
前記抽出された複合語の前記検索キー文書および前記検索対象文書での出現頻度を算出することを特徴とする類似文書検索方法。

【請求項5】 ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索するためのプログラムが記録された記録媒体であって、
前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、
前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、
前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語数の上限値を指定する指定手段と、
前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の単語数で構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、
前記複合語抽出手段によって抽出された複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備するプログラムが記録されていることを特徴とする記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文書データベースから、文書間の類似度に基づく文書データの検索を行う類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体に関する。

【0002】

【従来の技術】近年、大量の電子化された文書データが流通するようになり、自動分類等を行う目的で、文書データベース中から指定された文書（以下、検索キー文書と呼ぶ）に類似する文書の自動検索を行うシステムが実用化されてきている。この文書検索システムでは、検索キー文書に含まれている単語と検索対象となる文書（以下、検索対象文書と呼ぶ）に含まれている単語とを比較し、共通する単語の種類、出現場所、出現回数などから空間ベクトル法により類似度を算出して、類似度の高い検索対象文書を検索結果として出力する。

【0003】このような類似文書検索では、検索キー文書や検索対象文書から、その文書の特徴付ける単語を抽出することが、精度の高い類似文書検索を行うために非常に重要な鍵となる。これまで、文書から名詞やサ変名詞などの単語を対象に単語の抽出を行っていたが、文書から抽出された単一の単語が必ずしもその文書の特徴付ける単語として使用されているとは限らない。

【0004】そこで、文書の特徴付ける複数の単語の結合からなる複合語を文書から抽出する方法が提案されている。このように文書の特徴付ける複合語を抽出することで、単位の単語を抽出する方法に比べ、ユーザの意図する検索結果をより高い精度で得ることができる。

【0005】

【発明が解決しようとする課題】ところで、このように複合語を抽出する方法では、複合語の語長（単語数）が長くなればなるほど、その複合語がその文書の特徴付ける度合も高くなるが、その反面、検索対象の文書から抽出される単語の数が極端に減ってしまう傾向がある。このため、本来類似文書として検索されるべき文書に対する類似度として妥当な値が得られず、やはりユーザの意図する検索結果を高精度に得ることは困難であった。

【0006】本発明は、このような課題を解決するためのもので、検索キー文書と検索対象文書から複合語を抽出して類似文書を検索する場合の検索精度の向上を図ることのできる類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体の提供を目的とする。

【0007】

【課題を解決するための手段】上記した目的を達成するために、本発明は、請求項1に記載されるように、ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語数の上限値を指定する指定手段と、前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の数の単語により構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、前記複合語抽出手段によって抽出された複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備することを特徴とする。

【0008】本発明の類似文書検索装置では、文書から抽出された単語の結合により構成される複合語のうち、指定された上限値以下の数の単語により構成されるすべての複合語を検索キー文書および検索対象文書から抽出

することで、文書の特徴付ける複合語を漏れなく抽出することができ、検索キー文書と検索対象文書との類似度をより高精度に算出することができる。

【0009】また、本発明は、請求項2に記載されるように、ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語の数の上限値を指定する指定手段と、前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の数の単語により構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、任意の単語を不要語として選択する不要語選択手段と、前記複合語抽出手段によって抽出された複合語のうち前記不要語選択手段によって選択された不要語を含む複合語を無効とする複合語無効化手段と、前記複合語抽出手段によって抽出された有効な複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備することを特徴とする。

【0010】本発明の類似文書検索装置では、文書の特徴付ける複合語を漏れなく抽出することができ、検索キー文書と検索対象文書との類似度をより高精度に算出することができるとともに、予め指定された不要語の単語を含む複合語を無効なものとするすることで、ユーザの意図をさらに反映したより一層高精度な類似文書検索を行うことが可能になる。

【0011】

【発明の実施の形態】以下、図面を参照して本発明の実施形態を説明する。

【0012】図1は本発明の実施形態である類似文書検索装置のハードウェア構成を示すブロック図である。

【0013】同図に示すように、この類似文書検索装置は、キーボードなどの入力装置1、CPUおよびメモリなどから構成される制御装置2、類似文書の検索結果などを表示する表示装置3、および、文書データや類似文書検索のための各文書の単語情報や品詞辞書、不要語辞書などを格納する外部記憶装置4などから構成されている。外部記憶装置4に格納された品詞辞書、不要語辞書の構成を図6、図7にそれぞれ示す。

【0014】図2にこの類似文書検索装置における制御装置1の構成を示す。制御装置1は、制御部100とメモリ部200から構成される。

【0015】制御部100は、初期化部101、入力部102、出力部103、抽出対象品詞設定部104、最大結合数設定部105、検索対象文書読み出し部10

6、検索対象単語切り出し部107、検索対象単語複合語抽出部108、検索対象単語出現頻度算出部109、検索対象単語情報書き込み部110、検索キー文書入力部111、検索キー単語切り出し部112、検索キー単語複合語抽出部113、検索キー単語出現頻度算出部114、検索対象単語情報読み出し部115、共通単語抽出部116、類似度算出部117、検索結果出力部118などから構成される。メモリ部200は、品詞情報バッファ部201、選択品詞情報バッファ部202、不要語情報バッファ部203、最大結合数バッファ部204、検索対象文書格納バッファ部205、検索対象全文分割単語格納バッファ部206、検索対象複合語候補格納バッファ部207、検索対象抽出複合語格納バッファ部208、検索対象単語情報格納バッファ部209、検索キー文書格納バッファ部210、検索キー全文分割単語格納バッファ部211、検索キー複合語候補格納バッファ部212、検索キー抽出複合語格納バッファ部213、検索キー単語情報格納バッファ部214、共通単語情報格納バッファ部215、算出類似度格納バッファ部216、検索結果出力バッファ部217などから構成される。

【0016】初期化部101は、上記各バッファ部の初期化を行い、更に、外部記憶装置4における辞書（品詞辞書、不要語辞書など）の内容をメモリ部に読み込む。

【0017】入力部102は、ユーザによる入力装置1からの検索キー文書や単語抽出条件の設定など各種設定の入力を行う。

【0018】出力部103は、入力部102により入力された検索キー文書などの各種設定内容を表示装置3に出力する。

【0019】抽出対象品詞設定部104は、ユーザが品詞情報バッファ部201から選択した抽出対象単語の品詞を選択品詞情報バッファ部202に格納する。

【0020】最大結合数設定部104は、ユーザが指定した複合語の最大結合数を最大結合数バッファ部204に格納する。

【0021】検索対象文書読み出し部106は、外部記憶装置4に格納されている検索対象文書に関する情報を文書データベース化するために、文書データベース化すべき文書情報を外部記憶装置4から読み込み、検索対象文書格納バッファ部205に格納する。

【0022】検索対象単語切り出し部107は、検索対象文書格納バッファ部205に格納されている検索対象文書からの単語切り出しを行う。そして、その検索対象文書から抽出される全ての単語とその品詞を検索対象文書全文分割単語格納バッファ部206に格納する。単語の切り出しは形態素解析などにより行い、その文書から抽出される単語の品詞情報を「名詞」、「サ変名詞」、「形容詞」などで表現する。

【0023】検索対象単語複合語抽出部108は、検索

対象全文分割単語格納バッファ部206に格納されている全ての単語とその品詞情報の中から、選択品詞情報バッファ部202に格納されている品詞情報を参照して、該当する単語群（1つ、または複数の単語）を順次抽出し、検索対象複合語候補格納バッファ部207に格納する。さらに、検索対象複合語候補バッファ部207に格納されている単語群から最大結合数バッファ部204に格納されている結合数以下の複合語を抽出し、検索対象抽出複合語格納バッファ部208に格納する。

【0024】検索対象単語出現頻度算出部109は、検索対象単語複合語抽出部208により抽出された個々の単語や複合語について、検索対象文書中での出現頻度を算出し、これを検索対象文書の単語情報として検索対象単語情報格納バッファ部209に格納する。

【0025】検索対象単語情報書き込み部110は、検索対象単語情報格納バッファ部209に格納されている検索対象文書の単語情報を外部記憶装置4に格納する。

【0026】検索キー文書入力部111は、入力装置1から入力された検索キー文書の情報を検索キー文書格納バッファ部210に格納する。

【0027】検索キー単語切り出し部112は、検索キー文書格納バッファ部210に格納されている検索キー文書からの単語切り出しを行う。そして、その検索キー文書から抽出される全ての単語とその品詞を検索キー文書全文分割単語格納バッファ部211に格納する。単語の切り出しは形態素解析などにより行い、その文書から抽出される単語の品詞情報を「名詞」、「サ変名詞」、「形容詞」などで表現する。検索キー単語複合語抽出部113は、検索キー全文分割単語格納バッファ部211に格納されている全ての単語とその品詞情報の中から、選択品詞情報バッファ部202に格納されている品詞情報を参照して、該当する単語群（1つ、または複数の単語）を順次抽出し、検索キー複合語候補格納バッファ部212に格納する。さらに、検索キー複合語候補格納バッファ部212に格納されている単語群から最大結合数バッファ部に格納されている結合数以下の複合語を抽出し、検索キー抽出複合語格納バッファ部213に格納する。

【0028】検索キー単語出現頻度算出部114は、検索キー単語複合語抽出部113により抽出された個々の単語や複合語について、検索キー文書中での出現頻度を算出し、これを検索キー文書の単語情報として検索対象単語情報格納バッファ部214に格納する。

【0029】検索対象単語情報読み出し部115は、外部記憶装置4に格納されている各検索対象文書の単語情報（単語の出現頻度情報）を1文書毎に呼び出し、検索対象単語情報格納バッファ部209に格納する。

【0030】共通単語抽出部116は、検索キー単語情報格納バッファ部214に格納されている検索キー文書の単語情報と検索対象単語情報格納バッファ部209に

格納されている検索対象文書の単語情報とを比較して、一致する単語の種類と出現頻度情報を共通単語情報格納バッファ部215に格納する。

【0031】類似度算出部117は、共通単語情報格納バッファ部215に格納されている情報に基づき、検索キー文書と検索対象文書との類似度を算出し、その類似度を算出類似度格納バッファ部216に格納する。

【0032】検索結果出力部118は、算出類似度格納バッファ部216に格納されている各検索対象文書の類似度値を適宜並べ替えて、検索結果出力バッファ部217に格納し、さらに検索結果出力バッファ部217の内容を表示装置3に出力する。次に、本実施形態の類似文書検索装置の動作を説明する。

【0033】最初に検索対象文書データベースの作成手順を図3のフローチャートにより説明する。

【0034】まず、初期化部101により全メモリ部の初期化を行い、外部記憶装置4の品詞辞書と不要語辞書の情報をそれぞれ品詞情報バッファ部201、不要語情報バッファ部203に格納する（ステップ301）。品詞情報バッファ部201の構成を図9に、不要語情報バッファ部203の構成を図11にそれぞれ示す。

【0035】続いて抽出対象品詞設定部104が起動され、入力装置1を通じてユーザより抽出する単語の品詞選択の入力を受け付けて、図10に示すように、抽出対象品詞を選択品詞情報バッファ部202に格納する。また、同様に最大結合数設定部105が起動され、入力装置1を通じてユーザより抽出する複合語の最大結合数値の入力を受け付けて、図12に示すように、最大結合数バッファ部204に格納する（ステップ302）。

【0036】これらの設定が完了すると、検索対象文書読み出し部106が外部記憶装置4から複数のテキスト文書を読み出し、検索対象文書格納バッファ部205に検索対象文書として格納する（ステップ303）。具体例として、例えば、図13に示すような内容のテキスト文書を検索対象文書として格納されたとする。

【0037】次に、検索対象単語切り出し部107が、検索対象文書格納バッファ部205に格納されている検索対象文書について、形態素解析などによって単語の切り出しを行い、切り出した単語とその品詞情報を、図14に示すように、検索対象全文分割単語格納バッファ部206に格納する（ステップ304）。

【0038】続いて検索対象単語複合語抽出部108が起動される。検索対象単語複合語抽出部108は、検索対象全文分割単語格納バッファ部206に格納されている当該検索対象文書の全単語とその品詞情報、選択品詞情報バッファ部202を参照し、該当する単語（1個以上）を、図15に示すように、検索対象複合語候補格納バッファ部207に格納する。

【0039】さらに、検索対象単語複合語抽出部108は、検索対象複合語候補格納バッファ部207に格納さ

れている1個、または複数の単語群、不要語情報バッファ部203の不要語情報、そして最大結合数バッファ部204の最大結合数値を参照し、1以上かつ最大結合数以下の単語の結合からなる複合語を検索対象文書の中から抽出し、図16に示すように、検索対象抽出複合語格納バッファ部208に全て格納する（ステップ305）。なお、ここで抽出される複合語には、単独の単語、つまり結合数1の単語も含むものとする。

【0040】ここで、検索対象単語複合語抽出部108により、単語、複合語が少なくとも1個以上抽出された場合（ステップ306）、検索対象単語出現頻度算出部109が起動される。検索対象単語出現頻度算出部109は、検索対象抽出複合語格納バッファ部208に格納されている複合語について、当該検索対象文書中での出現頻度を複合語別に累積し、図17に示すように、検索対象単語情報格納バッファ部209に順次格納する（ステップ307）。検索対象語情報格納バッファ部209において、複合語（単語も含む）と頻度とは対応して登録されており、例えば、単語「住所録」は当該文書中に4回出現していることを表す。

【0041】以上の複合語抽出処理と抽出単語出現頻度算出処理は当該文書の文末まで行われる（ステップ308）。

【0042】当該文書の複合語抽出処理と抽出単語出現頻度算出処理が終了すると、図17に示す検索対象単語情報格納バッファ部209に格納された情報は、検索対象文書のデータベースとして外部記憶装置4に蓄積される（ステップ309）。

【0043】これで1検索対象文書のデータベースへの蓄積が終了するが、検索対象文書格納バッファ部205にまだ検索対象文書が残っている場合、ステップ304にもどって、前記同様の文書データベース生成が行われる。検索対象文書が残っていない場合、データベースの生成は終了する。

【0044】ここで、検索対象単語複合語抽出部108による単語複合語抽出（ステップ305）の手順を図5、図6、図14～図16を使って詳しく説明する。

【0045】まず、初期化として現結合数に0を代入する（ステップ510）。複合語候補となる結合単語を抽出するため、図14に示す検索対象全文分割単語格納バッファ部206に記憶されている単語に対応する品詞情報と、図10に示す選択品詞情報バッファ部202にある品詞情報とを比較し（ステップ502）、検索対象全文分割単語格納バッファ部206に記憶されている単語が対象品詞の単語であった場合、図15に示すように、当該単語を検索対象複合語候補格納バッファ部207に格納し（ステップ503）、現結合数に1を加える（ステップ504）。対象品詞の単語をすべて抽出したらステップ505に移る。

【0046】ステップ505では現結合数を調べ、現結

合数が0より大きかった場合は処理を続行し、0であった場合は複合語抽出処理を終了する。

【0047】現結合数が0より大きい場合は、続いて、複合語抽出のための先頭カウンタと結合数カウンタにそれぞれ1をセットして初期化を行う（ステップ506）。

【0048】ここから、複合語の抽出が、結合数カウンタが最大結合数バッファ部204が示す値になるまで以下のように行われる（ステップ507）。

【0049】先頭カウンタが示す検索対象複合語候補格納バッファ部207の単語から結合数カウンタの示す単語数の複合語を抽出できる場合（ステップ508）、ステップ509からステップ512にかけて不要語チェックを行う。不要語チェックは、当該複合語を構成する単語と図11に示す不要語情報バッファ部203の単語とを全て比較し、不要語に該当するものがあった場合、その不要語を含む複合語を抽出の対象としない処理を行う（ステップ513）。

【0050】なお、ここでは複合語の一要素となる単語が不要語であった場合、複合語抽出の対象としない処理を行ったが、そうした不要語が複合語の頭に接頭する、あるいは、末尾に接尾する場合にだけ、複合語抽出の対象としない処理を行うようにしてもよい。

【0051】不要語にあたる単語が、抽出された複合語に含まれない場合は、その複合語を検索対象抽出複合語格納バッファ部208に格納する（ステップ514）。

【0052】このときの結合数カウンタが示す単語数の複合語を図15に示す検索対象複合語候補格納バッファ部207からすべて抽出する（ステップ514）。すべて抽出したら、結合数カウンタに1を加え（ステップ515）、先頭から新たな結合数カウンタが示す結合数の複合語の抽出を行う（ステップ516）。

【0053】そして結合数カウンタが最大結合数バッファ部204が示す値を超えた場合、複合語抽出を終了する（ステップ507）。

【0054】次に、類似文書の検索手順を、図4のフローチャートにより説明する。

【0055】まず、初期化部101により全メモリ部を初期化し、外部記憶装置4の品詞辞書と不要語辞書の情報をそれぞれ品詞情報バッファ部201、不要語情報バッファ部203に格納する。（ステップ401）。続いて抽出対象品詞設定部104が起動され、入力装置1を通じてユーザより抽出する単語の品詞選択の入力を受け付けて抽出対象品詞を選択品詞情報バッファ部202に格納する。また、同様に最大結合数設定部105が起動され、入力装置1を通じてユーザより抽出する複合語の最大結合数値の入力を受け付けて最大結合数バッファ部204に格納する（ステップ402）。

【0056】続いて、検索キー文書入力部111が起動され、入力装置1を通じてユーザより検索キーとなる文

書の入力を受け付けて検索キー文書格納バッファ部210に格納する（ステップ403）。具体例として、例えば、図18に示すような内容のテキスト文書を検索キー文書として格納したとする。

【0057】次に、検索キー単語切り出し部112が、検索キー文書格納バッファ部210に格納されている検索キー文書について、形態素解析などによって単語の切り出しを行い、切り出した単語とその品詞情報を検索キー全文分割単語格納バッファ部211に格納する（ステップ404）。

【0058】そして、検索キー単語複合語抽出部113が起動される。検索キー単語複合語抽出部113は、検索キー全文分割単語格納バッファ部211に格納されている当該検索キー文書の全単語とその品詞情報、選択品詞情報バッファ部201、202を参照し、該当する単語（1個以上）を検索キー複合語候補格納バッファ部212に格納する。さらに、検索キー単語複合語抽出部113は、検索キー複合語候補格納バッファ部212に格納されている1個、または複数の単語群、不要語情報バッファ部203の不要語情報、そして最大結合数バッファ部204の最大結合数値を参照し、1以上かつ最大結合数以下の単語が結合した複合語を検索キー文書の中から抽出し、これらを検索キー抽出複合語格納バッファ部213に全て格納する（ステップ405）。

【0059】なお、ここで抽出される複合語には、単独の単語、つまり結合数1の単語も含むものとする。

【0060】ここで、検索キー単語複合語抽出部113により、単語、複合語が少なくとも1個以上抽出された場合（ステップ406）、検索キー単語出現頻度算出部114が起動される。検索キー単語出現頻度算出部114は、検索キー抽出複合語格納バッファ部213に格納されている複合語について、当該検索キー文書中での出現頻度を複合語別に累積し、検索キー単語情報格納バッファ部214に順次格納する（ステップ407）。図22に検索キー単語情報格納バッファ部214の格納例を示す。この検索キー単語情報格納バッファ部214において、複合語（単語も含む）と頻度は対応しており、例えば、単語「筆文字」は当該文書中に3回出現していることを表す。

【0061】この複合語抽出処理と抽出単語出現頻度算出処理を当該文書の文末まで行う（ステップ408）。これで検索キー文書の単語情報の生成が終了する。

【0062】次に、検索対象単語情報読み出し部115が、外部記憶装置4に格納されている各検索対象文書の単語情報を1文書毎に読み込み、検索対象単語情報格納バッファ部209に格納する（ステップ409）。

【0063】続いて、共通単語抽出部116が起動され、検索対象単語情報格納バッファ部206と検索キー単語情報格納バッファ部214とに共通して格納されている単語、複合語を共通単語情報格納バッファ部215

に格納する。具体例を図23に示す。図17の検索対象単語情報格納バッファ部209と図22の検索キー単語情報格納バッファ部214に共通する単語（複合語）として、「宛名印刷」が抽出され、この「宛名印刷」とその頻度が検索キー側、検索対象側それぞれ1、2というように対応づけて格納する（ステップ410）。

【0064】次に、類似度算出部117が、共通単語情報格納バッファ部215に格納されている頻度情報に基づき検索キーと検索対象文書との類似度を空間ベクトル法などにより算出し、その類似度値を算出類似度格納バッファ部216に格納する（ステップ411）。例えば、図24に示すように、各検索対象文書ごとの類似度が算出類似度格納バッファ部216に格納される。

【0065】全ての検索対象文書について類似度計算が終了すると（ステップ412）、検索結果出力部118は、算出類似度格納バッファ部216に格納されている各検索対象文書ごとの類似度を類似度が高い順に並べ替えて検索結果出力バッファ部217に格納し、そのバッファの内容を表示装置3に出力する。出力結果は、例えば、図26が示すような形で出力される（ステップ413）。なお、図26では類似度値に閾値を設けて表示しているが、類似度値に一定の閾値を設けて、検索結果として表示する検索対象文書の量を制限できるようにしてもよい。

【0066】これで1検索キー文書の類似文書検索は終了するが、新たに検索キー文書がある場合、ステップ402に戻って、同様な処理を行う。検索キー文書がなければ検索処理はこれで終了する（ステップ414）。

【0067】なお、この類似文書検索処理における単語複合語抽出の手順は、検索対象文書データベース作成における処理と対象となる文書による処理部、バッファ部の違いはあるが全く同様の処理である。

【0068】以上のように、本実施形態の類似文書検索装置では、検索キー文書あるいは検索対象文書に、例えば「筆文字宛名印刷機能」といった複合語が存在するならば、最大結合数を3としたとき「筆文字宛名」「文字宛名印刷」「宛名印刷機能」「筆文字」「文字宛名」「宛名印刷」「印刷機能」などの各所において部分的に連続した複合語が抽出される。このように、特定の内容の文書の特徴付ける異なる複合語を漏れなく抽出することができるので、文書間のより妥当な類似度を計算でき、ユーザの意図する精度の高い類似文書検索を行うことができる。また、検索対象となる文書全体から抽出される単語種の総数が少なくなり、データベースの規模を縮小することができる。

【0069】さらに、本実施形態の類似文書検索装置では、文書中から複合語を抽出するとき、ユーザにより指定された不要語の単語を含む複合語を無効なものとするので、ユーザの意図する類似文書検索をさらに高精度に行うことが可能になる。

【0070】なお、以上説明した類似文書検索装置は、例えば、汎用的なハードウェア環境に、フロッピーディスク、CD-ROMなどの記録媒体に記録されたアプリケーションプログラムを追加することによっても提供することが可能である。

【0071】

【発明の効果】以上説明したように、本発明によれば、文書から抽出された単語の結合により構成される複合語のうち、指定された上限値以下の数の単語により構成されるすべての複合語を検索キー文書および検索対象文書から抽出することで、文書の特徴付ける複合語を漏れなく抽出することができ、検索キー文書と検索対象文書との類似度をより高精度に算出することができる。また、予め指定された不要語の単語を含む複合語を無効なものとするので、ユーザの意図をさらに反映したより一層高精度な類似文書検索を行うことが可能になる。

【図面の簡単な説明】

【図1】本発明の実施形態である類似文書検索装置のハードウェア構成を示すブロック図

【図2】図1の制御装置の内部構成を示すブロック図

【図3】本実施形態の類似文書検索装置の検索対象文書データベース生成の動作手順を示すフローチャート

【図4】本実施形態の類似文書検索装置の類似文書検索の動作手順を示すフローチャート

【図5】複合語抽出処理の詳細な動作手順を示すフローチャート

【図6】図5と同じく複合語抽出処理の詳細な動作手順を示すフローチャート

【図7】品詞辞書の内容を示す図

【図8】不要語辞書の内容を示す図

【図9】品詞情報バッファの内容を示す図

【図10】選択品詞情報バッファの内容を示す図

【図11】不要語情報バッファの内容を示す図

【図12】最大結合数バッファの内容を示す図

【図13】検索対象文書格納バッファの内容を示す図

【図14】検索対象全文分割単語格納バッファの内容を示す図

【図15】検索対象複合語候補格納バッファの内容を示す図

【図16】検索対象抽出複合語格納バッファの内容を示す図

【図17】検索対象単語情報格納バッファの内容を示す図

【図18】検索キー文書格納バッファの内容を示す図

【図19】検索キー全文分割単語格納バッファの内容を示す図

【図20】検索キー複合語候補格納バッファの内容を示す図

【図21】検索キー抽出複合語格納バッファの内容を示す図

【図22】 検索キー単語情報格納バッファの内容を示す図

【図23】 共通単語情報格納バッファの内容を示す図

【図24】 算出類似度格納バッファの内容を示す図

【図25】 検索結果出力バッファの内容を示す図

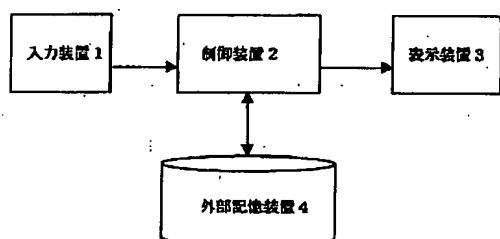
【図26】 類似文書検索結果の出力例を示す図

【符号の説明】

- 100 制御部
101 初期化部
102 入力部
103 出力部
104 抽出対象品詞設定部
105 最大結合数設定部
106 検索対象文書読み出し部
107 検索対象単語切り出し部
108 検索対象単語複合語抽出部
109 検索対象単語出現頻度算出部
110 検索対象単語情報書き込み部
111 検索キー文書入力部
112 検索キー単語切り出し部
113 検索キー単語複合語抽出部
114 検索キー単語出現頻度算出部

- 115 検索対象単語情報読み出し部
116 共通単語抽出部
117 類似度算出部
118 検索結果出力部
200 メモリ部
201 品詞情報バッファ部
202 選択品詞情報バッファ部
203 不要語情報バッファ部
204 最大結合数バッファ部
205 検索対象文書格納バッファ部
206 検索対象全文分割単語格納バッファ部
207 検索対象複合語候補格納バッファ部
208 検索対象抽出複合語格納バッファ部
209 検索対象単語情報格納バッファ部
210 検索キー文書格納バッファ部
211 検索キー全文分割単語格納バッファ部
212 検索キー複合語候補格納バッファ部
213 検索キー抽出複合語格納バッファ部
214 検索キー単語情報格納バッファ部
215 共通単語情報格納バッファ部
216 算出類似度格納バッファ部
217 検索結果出力バッファ部

【図1】



【図11】

不要語情報バッファ

下記↓
検索↓
辞書↓
上記↓
上記↓
↓

(↓:文字終端)

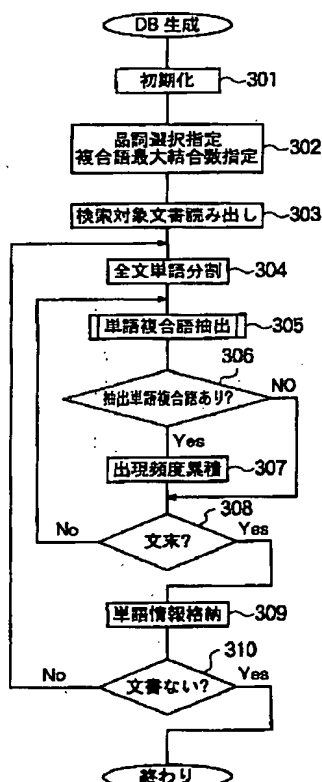
【図13】

検索対象文書格納バッファ

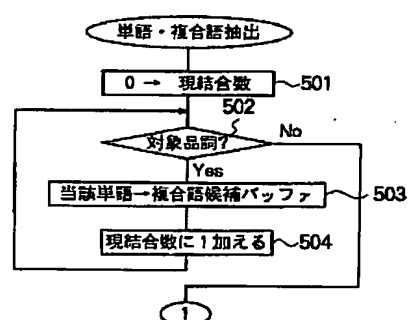
無文字宛名印刷機能を中心とするソフトで、住所録データベースのデータを変換するためのツールが.....

(↓:文字終端)

【図3】



【図5】



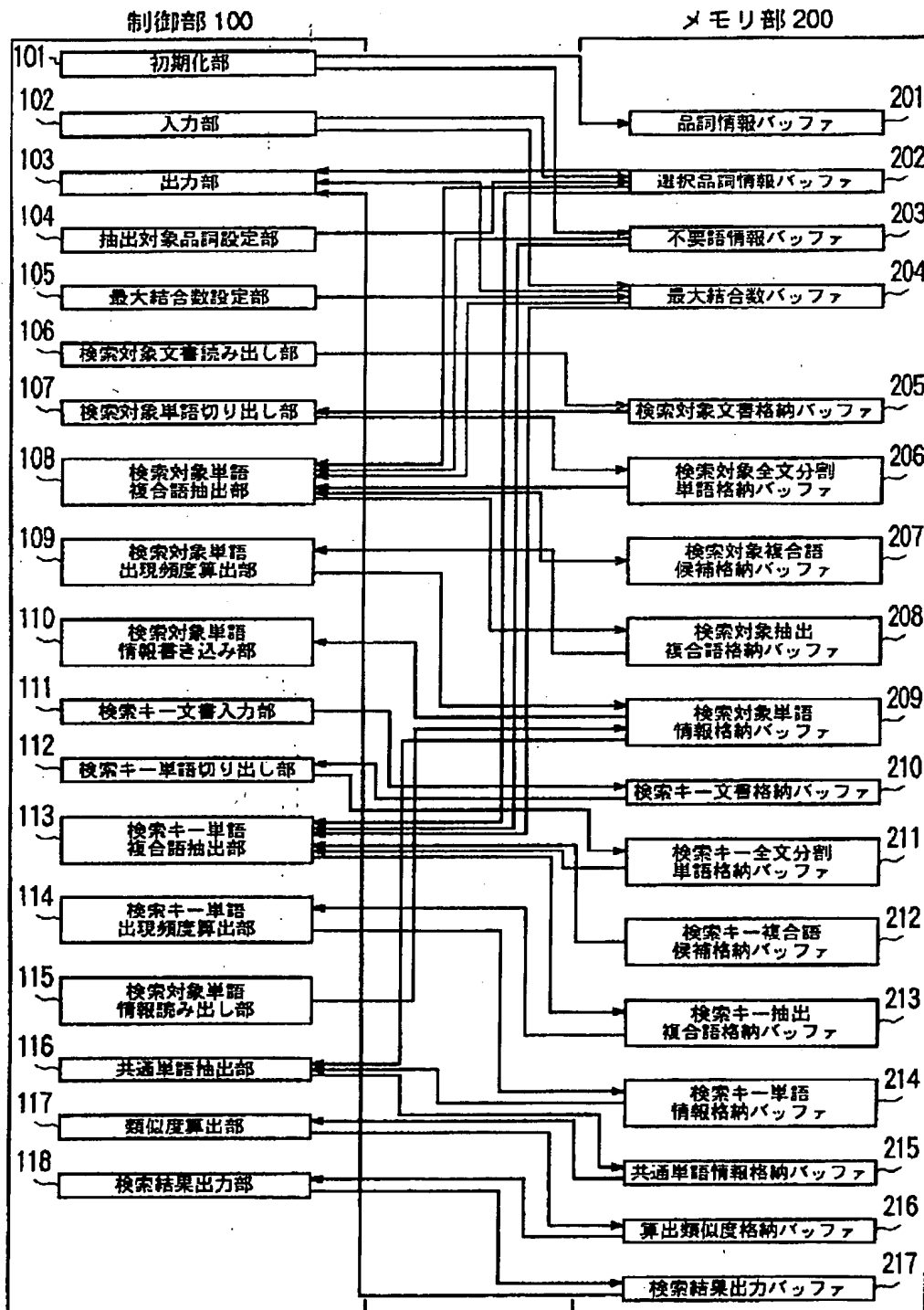
【図19】

検索キー全文分割単語格納バッファ

筆↓	名詞↓
文字↓	名詞↓
に↓	格助詞↓
:	:
上記↓	名詞↓
宛名↓	名詞↓
印刷↓	サ変名詞↓
や↓	格助詞↓
デジカメ↓	名詞↓
:	:

(↓:文字終端)

【図2】



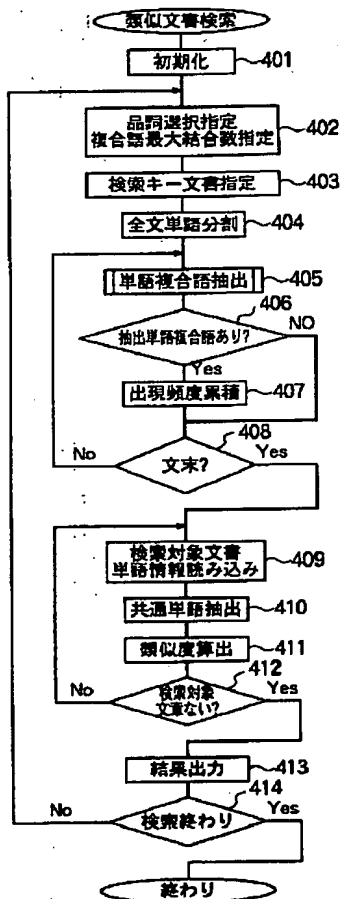
【図12】

最大結合数バッファ

31

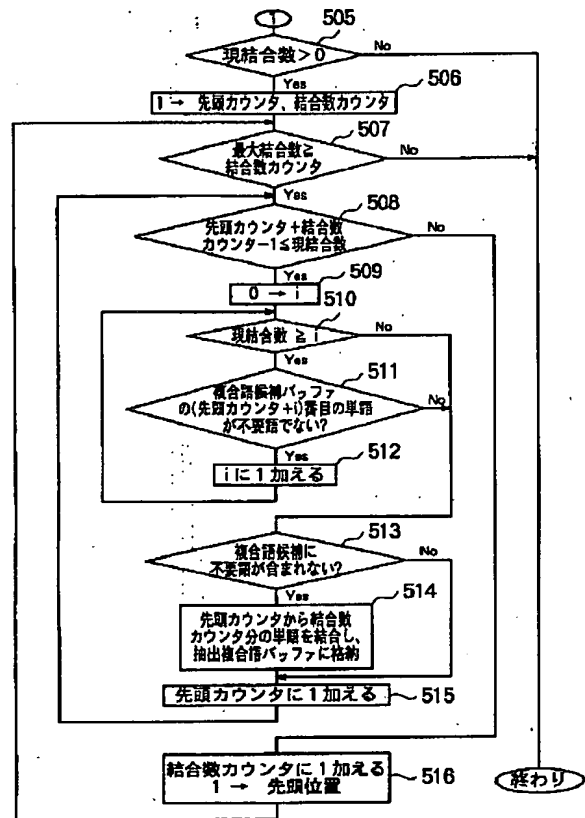
(1: 文字列)

【図 4】



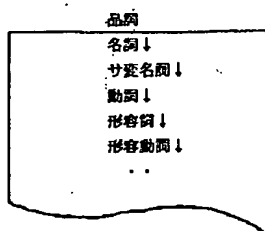
【図 7】

【図 6】



【図 8】

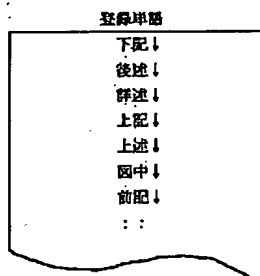
品詞辞書



(↓: 文字終端)

【図 10】

不要語辞書



(↓: 文字終端)

【図 15】

選択品詞情報バッファ

名詞↓
サ変名詞↓
形 詞↓
形容動詞↓

(↓: 文字終端)

検索対象複合語候補 納バッファ

筆↓	名詞↓
文字↓	名詞↓
宛名↓	名詞↓
印刷↓	サ変名詞↓
機能↓	名詞↓

(↓: 文字終端)

【図9】

品詞情報バッファ

名詞↓
サ変名詞↓
形容詞↓
形容動詞↓
：
副詞↓
格助詞↓

(↓:文字終端)

【図14】

検索対象全文分割単語格納バッファ

筆↓	名詞↓
文字↓	名詞↓
宛名↓	名詞↓
印刷↓	サ変名詞↓
機能↓	名詞↓
を↓	格助詞↓
中心↓	名詞↓
と↓	格助詞↓
：	：

(↓:文字終端)

【図16】

【図17】

検索対象抽出複合語格納バッファ

筆文字宛名↓
文字宛名印刷↓
：
宛名印刷機能↓
：
宛名印刷↓
：

(↓:文字終端)

検索対象単語情報格納バッファ

単語	頻度
筆文字宛名↓	2↓
宛名印刷↓	2↓
：	：
住所録↓	4↓
：	：
変換↓	8↓
：	：

(↓:文字終端)

【図18】

【図20】

【図21】

検索キー文書格納バッファ

筆文字フォントによる上記宛名印刷やデジ
カメ画像からの映像取り込み、住所録のデ
ータベースのインポート/エクスポート機
能が充実..... ↓

(↓:文字終端)

検索キー複合語候補格納バッファ

上記↓	名詞↓
宛名↓	名詞↓
印刷↓	サ変名詞↓

(↓:文字終端)

検索キー抽出複合語格納バッファ

宛名印刷↓
宛名↓
印刷↓

(↓:文字終端)

【図22】

【図23】

【図25】

検索キー単語情報格納バッファ

単語	頻度
筆文字↓	3↓
宛名印刷↓	1↓
：	：
住所録↓	6↓
：	：
機能↓	4↓

共通単語情報格納バッファ

単語	検索キー側頻度	検索対象側頻度
宛名印刷↓	1↓	2↓
住所録↓	6↓	4↓
変換↓	8↓	3↓
機能↓	4↓	1↓
：	：	：

(↓:文字終端)

検索結果出力バッファ

類似検索対象文書
3↓
51↓
289↓
：

(↓:文字終端)

【図24】

【図26】

算出類似度格納バッファ

検索対象文書番号	類似度
1↓	0.03198↓
2↓	0.16771↓
3↓	0.49534↓
：	：

(↓:文字終端)

検索結果出力例

類似文書検索結果

<文書番号>
3
51
289
:

フロントページの続き

(72)発明者 仁科 卓哉
東京都青梅市新町1381番地1 東芝コンピ
ュータエンジニアリング株式会社内

(72)発明者 中本 幸夫
東京都青梅市新町1381番地1 東芝コンピ
ュータエンジニアリング株式会社内